Design and Evaluation of Multimedia Stimuli to Evoke Clinical Concepts

Jeremy C. Wyatt[1], William M. Detmer[2] and Lawrence M. Fagan[2]
[1]Biomedical Informatics Unit, Imperial Cancer Research Fund, London WC2A 3PX, UK
[2]Section on Medical Informatics, Stanford University Medical School, Stanford CA 94305-5479

Continuous speech recognition systems have the potential to facilitate clinical data entry, but evaluating them rigorously is difficult. We describe a tool to aid evaluators of such systems. The tool is a HyperCard stack with stimuli consisting of pictures, sounds and the minimum of words to evoke 20 QMR physical findings. Despite using up to four different stimuli to communicate each finding and piloting the material on six subjects, eight test subjects made a total of 66 errors (42%) in interpreting the 20 sets of stimuli, of which 22 errors (14%) were serious. These results are relevant to those designing interfaces for decision-support, tutorial and student testing systems.

INTRODUCTION

One factor limiting the acceptability of clinical information systems is the time and effort required to input patient data. Continuous speech recognition interfaces have the potential to reduce this, and several have been built [eg. 1]. Ideally, the usability and accuracy of such systems should be evaluated in real-world settings, where users are free to communicate naturally with the computer. However, this is not always practical. We therefore assembled a multimedia resource to simulate a real-world setting, so that we could evaluate a speech interface to Quick Medical Reference (QMR) [2]. We needed to evoke in subjects' minds the concepts of 20 abnormal physical findings, which they would express to the computer in their own words. The ideal stimulus material would have been real patients with stable findings, but this was not feasible. We rejected giving subjects text descriptions of physical findings to read out to the speech interface as such stimuli would constrain the words they used, and over-estimate its accuracy compared to use in a clinical environment.

We therefore decided to depict the 20 physical findings using mainly pictures and sounds. To eliminate bias due to learning and fatigue, the stimulus material was presented to each subject in a random order. We considered using video to present both pictures and sounds, but the need for subjects to see stimuli in a random order led us to present them using a Claris HyperCard stack. This paper analyses the errors made by eight subjects interpreting our stimuli, postulates why these errors occurred and suggests how to improve the communication of clinical concepts in other computer applications, such as advisory, teaching or student assessment programs.

METHODS

Selection of the test findings

The developers of the speech input system gave the evaluator (JCW) a list of all the QMR terms describing abnormal physical findings localised in the head, neck, chest and abdomen that the speech input system could understand. The evaluator randomised the 116 findings then, starting with the first, selected the 20 findings he considered most feasible to communicate using pictures or sounds alone or in combination (Figure 1). Findings that he believed could not be reliably communicated, like abdomen flank heavy or neck muscle flaccid, were rejected.

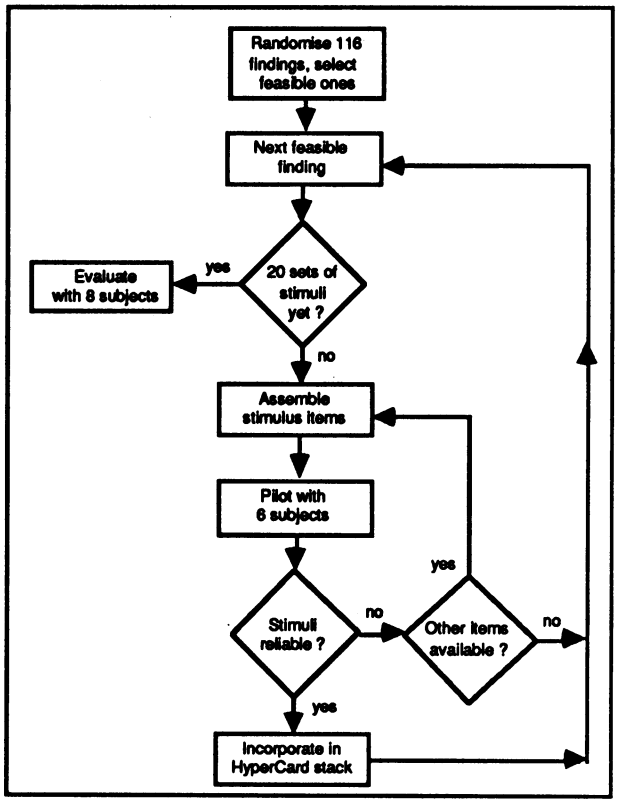

Figure 1. Procedure for selecting the findings and assembling and piloting the stimulus material

Table 1: The selected physical findings and the stimuli used to communicate them

QMR finding name	Text	Icon	Photo	Diagram	Sound	Total
abdomen flank bulging bilateral		specific	B & W			2
abdomen mass right lower quadrant	procedure			body chart		2
abdomen mass right upper quadrant	procedure			body chart		2
abdomen tenderness right lower quadrant		general		body chart	"Ouch !"	3
abdomen tenderness right upper quadrant				body chart	"Ouch !"	2
abdomen urinary bladder palpable or percussable	diagnosis & procedure	specific		body chart		4
artery carotid systolic bruit		specific	B & W		aortic stenosis murmur	3
bradycardia	"HR 45/min"	general				2
breast gynecomastia		specific	B & W			2
breast nipple retraction			B & W			1
breast tender		general	B & W		"Ouch !"	3
gallbladder palpable	diagnosis & procedure	specific		body chart		4
head and neck edema	diagnosis	general	color			3
heart murmur diastolic decrescendo second left interspace		specific	B & W	phono-cardiogram	mitral stenosis murmur	4
heart murmur systolic ejection second right interspace		specific	B & W	phono-cardiogram	aortic stenosis murmur	4
hepatomegaly present	diagnosis			diagram from book		2
splenomegaly massive	diagnosis	specific		body chart		3
splenomegaly moderate	diagnosis	specific		body chart		3
splenomegaly present	diagnosis	specific	B & W	body chart		4
umbilicus nodule			color	bodychart + arrow		2

Procedure for assembling the stimulus material
The evaluator scrutinised 11 books on physical examination, 3 medical atlases, 4 slide collections and tapes of cardiac murmurs for suitable stimuli to evoke the findings. If no suitable photograph could be found, he drew a diagram; most consisted of a standardised body chart with the outline of a mass (Figure 2) or shading to represent tenderness. Six pilot physician-subjects independently reviewed the test stimuli, blind to the finding, and were asked to name the abnormality. The stimuli were improved by selecting alternative or additional stimuli or, in one case, by deleting the finding and selecting the next feasible one from the randomised list.

In 14 of the 20 findings, pilot subjects found it difficult to gauge what level of detail was portrayed. For instance, a diagram showing splenomegaly moderate might be described as an abdominal mass, a left upper quadrant mass or splenomegaly. Icons were added to denote "*Be as specific as you can*" (10 findings) or "*This is a general finding*" (4 findings). The pilot studies showed that four physical finding could be evoked reliably by visual stimuli alone, while in 6 both visual and auditory stimuli were used.

In 9 of the remaining 10 findings text was added to describe either a typical patient in whom the finding might be observed (eg. for splenomegaly moderate: "Diagnosis: chronic malaria") or a procedure that would elicit it (eg. "Procedure: palpation"), being careful not to name the abnormality or site. In the remaining finding, bradycardia, a picture of an otherwise normal ECG with a rate 45 per minute failed to evoke the correct concept, so the words "HR: 45 per minute" and the "general" icon were used. Thus, 3 kinds of stimuli were used: visual items (photographs, diagrams or icons), sounds and text. Only one finding (breast nipple retraction) could be communicated reliably to pilot subjects by a single stimulus, and the mean number of stimuli per finding was 2.75 (see Table 1). Following piloting, we rejected one finding, heart impulse apical lateral displacement, from the list of QMR terms, as subjects could not guess it reliably even when a combination of various stimuli were used.

Implementation of the HyperCard stack
We built a HyperCard stack to display the stimulus material on a 13" high resolution color monitor. Two introductory cards gave instructions on how to use

the stack. Stimuli for each of the 20 findings were placed on one card. For each subject the order of the stimulus cards was randomised using a HyperCard function. Subjects navigated through the stack one card at a time using "Next Card" and "Previous Card" buttons; as they moved, the time and card name were logged to a file. Subjects could click a "Help" button at any time to remind them of the instructions.

The slide and photographs were scanned at a resolution which scaled to 72 dots per inch at the final screen size without aliasing. In some cases, the contrast or color balance were manipulated using Adobe Photoshop to improve the image. Where a suitable image could not be obtained or was inappropriate, a diagram was drawn and pasted onto the card. The images were saved as PICT files and displayed over the appropriate card using the "Picture" external command (XCMD). Two sounds from a tape of heart murmurs (one served for both aortic stenosis and carotid bruit) and a female voice saying "Ouch !" were digitised at 22kHz with 3:1 compression and incorporated into the stack as sound (SND) resources. A sample card, depicting gallbladder palpable, is shown in Figure 2.

Evaluation methods

Eight subjects were recruited from university hospitals. Selection criteria included being at least 2 years out of medical school and having no detailed knowledge of QMR terms nor speech interfaces. After a brief verbal introduction to the experiment, subjects read the two instruction cards and proceeded with the experiment. After viewing the stimuli on a card, the subject spoke a phrase or sentence describing the finding into a microphone. This utterance was captured by the speech interface and transcribed in real time by an investigator. Responses were classified using the following criteria:

- Correct response: location and finding correspond exactly to original QMR term
- Minor error: one anatomical location omitted (eg. facial edema instead of head & neck edema) or finding too specific (eg. tender right breast for breast tender)
- Major error: incorrect location, finding (eg. RUQ tenderness instead of RUQ mass) or degree of abnormality grossly over- or under-estimated (eg. splenomegaly for massive splenomegaly)

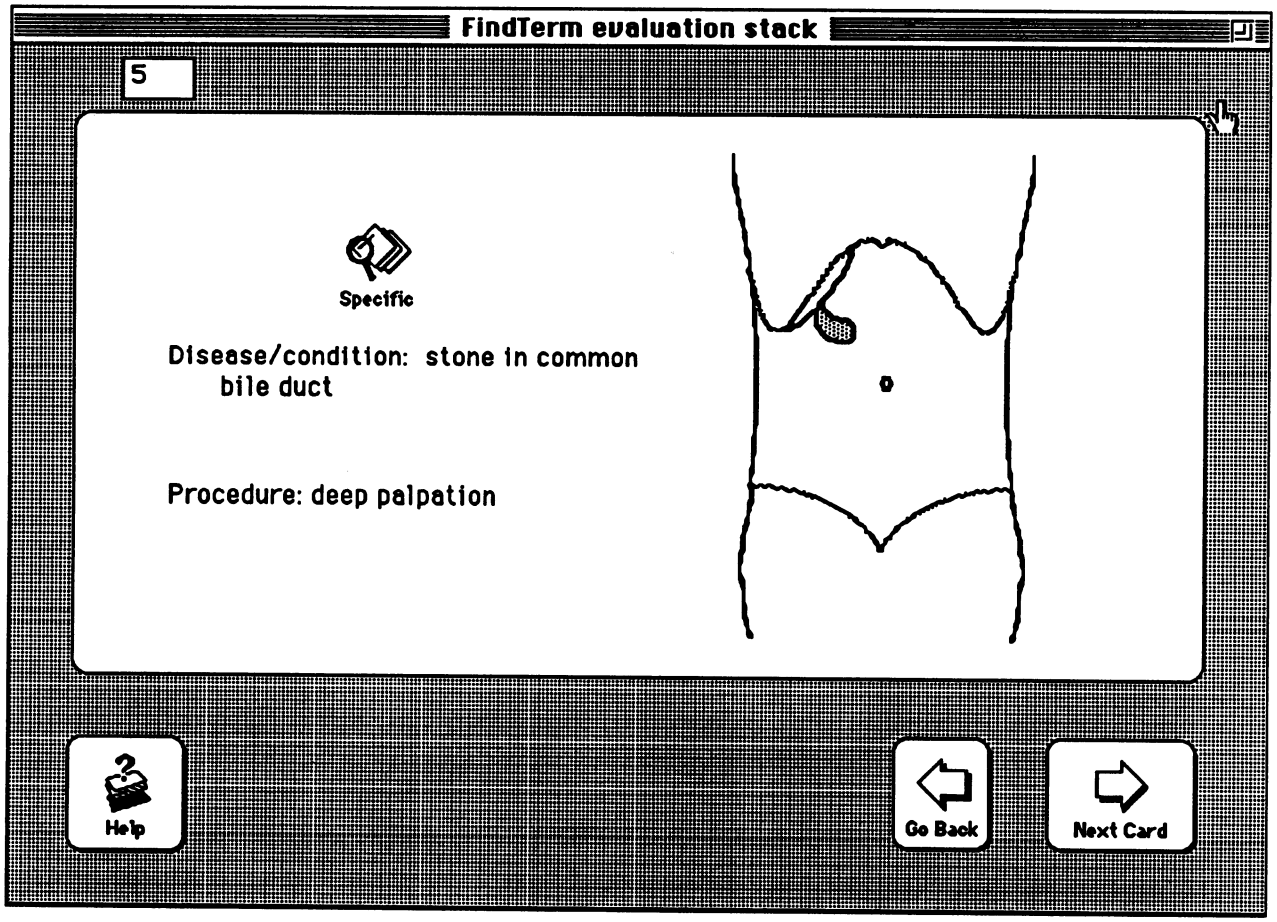

Figure 2. A sample card from the HyperCard stack depicting the finding gallbladder palpable

RESULTS

Of the 160 responses from 8 subjects, 94 (58%) were correct, 44 (28%) contained minor errors and 22 (14%) showed major errors.

Effect of subject factors on errors

For individual subjects, the error rate on 20 findings varied from 30% (2 major, 4 minor) to 65% (6 major, 7 minor). The error rate between individuals was not statistically significant (p=0.84, chi square test) nor was there a correlation between the error rate and the number of years since graduation from medical school (K = -0.08, p=0.84). A significant proportion (17, 77%) of the major errors occurred during the first half of each experiment (p=0.02, chi square) suggesting a training effect (see Figure 3).

Effect of finding type & number of stimuli on error rates

The major error rate per finding varied from 0/8 (for abdomen flank bulging bilateral, artery carotid bruit, bradycardia, breast gynecomastia, breast tender, murmur diastolic, splenomegaly present, splenomegaly moderate, umbilical nodule) to 5/8 for gallbladder palpable. The minor error rate varied from 0/8 (tender RLQ, bradycardia, gallbladder palpable, any splenomegaly) to 7/8 for breast tender, mainly because subjects were too specific about the site of tenderness.

There was a significant influence of the kind of finding (classed as a mass, tenderness, an abnormal sound or other) on the number of errors (p=0.048, chi square for any error vs. none), with errors occurring on 15 (63%) of 24 occasions for the three abnormal sounds. The body region (classed as head & neck, chest, abdomen) in which the finding was located had a significant influence on the errors (p=0.026, chi square for any error vs. none), with errors occurring on 11/16 (69%) occasions for the two findings located in the head and neck.

The total number of stimuli used to communicate the finding was not correlated with the total (p=0.3, least squares), minor (p=0.7) or major (p=0.33) error rates.

There was a significantly larger error rate in stimuli that did not include text (p=0.04, chi square) but no difference between those that did or did not include icons (p=0.65) or sounds (p=0.1).

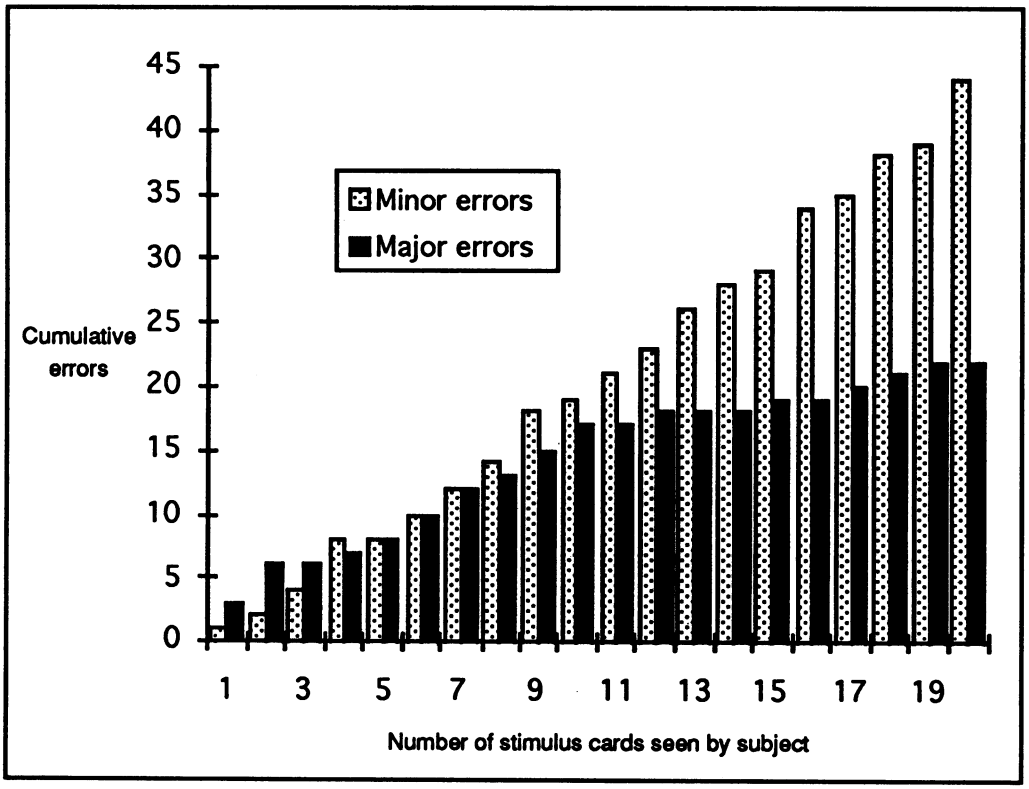

Figure 3. Graph of cumulative errors against number of stimulus cards seen by subjects

DISCUSSION

Finding suitable non-verbal stimulus material to communicate abnormal physical findings was a difficult exercise. We did not believe it could be done at all for 17 (46%) of 37 randomly selected findings, and needed to use text (alone or in combination) in 10 and icons to convey the correct level of detail in 14 of those findings we did attempt. The findings which we rejected were mainly those requiring tactile feedback; it is hard to see how these could be depicted using conventional methods, though "force-feedback" devices are now available for special purposes.

HyperCard appears to be a suitable medium for presenting stimulus material consisting of scanned photographs, sounds, text and diagrams; developers could also incorporate QuickTime movie clips, if relevant and available. The resolution of the images displayed is well within the 0.4mm required for a more exacting recognition task, interpreting chest radiographs [3].

Using a mean of 2.75 stimuli per finding, we were able to evoke the intended concept on a mean of 58% of occasions. It is interesting to compare this figure with the 55% agreement about which signs were present when physicians examined real patients with abnormal chest findings [4]. The number of stimuli per finding had no clear effect on the accuracy of evocation, perhaps because we deliberately used more stimuli to elicit complex concepts. We also deliberately added text stimuli where we thought subjects might experience difficulties; this appeared to reduce errors significantly.

Although we cannot yet be sure of the independent contribution of each factor to the error rate, some guidelines for evoking concepts in physicians' minds using the minimum of words are:

- Use pilot physicians to help select the material
- Give the test subjects plenty of practice
- Avoid certain body areas; our test physicians performed significantly worse on findings in the head and neck
- Avoid certain kinds of finding; our subjects made more errors with abnormal sounds.

Future work includes evaluating the responses of more physicians to the stimulus material and a multivariate analysis to quantify the independent contribution of each factor to the error rate.

These results highlight the difficulty of using pictorial material to communicate, and probably to teach, complex ideas, and are relevant to the design of user interfaces when clinical findings are shown on-screen to aid in decision-support [eg. 5], student testing or tutoring [6]. They also show that achieving a bias-free evaluation of a continuous speech recognition system, when speakers are not constrained to use a specific vocabulary, raises problems to which there is no easy solution. This is reminiscent of the evaluation of medical decision-aids [7].

Acknowledgements

We thank our subjects and Anne Brewer, Ramon Felciano, Chuck Friedman, Jay Heyman, Kevin Johnson, Chris Lane, Alex Poon and Smadar Shiffman for their help in developing software, designing and setting-up the experiments. JCW was in receipt of a UK Medical Research Council Travelling Fellowship at Stanford University. The speech-interface project was funded by the National Library of Medicine (grants LM-04864, LM-07033), and AHCPR (contract 213-89-0012). Computing resources were provided by the Stanford University CAMIS project, funded by the National Library of Medicine (grant LM-05305).

REFERENCES

1. Shiffman S, Wu A, Poon A et al. Building a speech interface to a medical diagnostic system. IEEE Expert 1991 (February): 41-50
2. Shiffman S, Lane C, Johnson K, Fagan L. The integration of a continuous-speech-recognition system with the QMR diagnostic program. Proc. 16th SCAMC. Washington, DC. November 1992.
3. Lams PM, Cocklin ML. Spatial resolution requirements for digital chest radiographs. Radiology 1986; 158: 11-19
4. Spiteri MA, Cook DG, Clarke SW (1988). Reliability of eliciting physical signs in examination of the chest. Lancet 1988; 1: 873-875
5. Nathwani B, Heckerman D, Horwitz E, Lincoln T. Integrated expert systems and videodisk in surgical pathology. Human Pathology 1990; 21: 11-27
6. Ingram D. Educational computing & medicine. In: Dalton K, Chard T (eds). Computers in Obstetrics & Gynaecology. Amsterdam: Elsevier 1990:313-28
7. Wyatt J, Spiegelhalter D. Field trials of medical decision-aids: potential problems and solutions. In Clayton P (ed). Proc. 15th SCAMC, Washington DC 1991. New York: McGraw Hill Inc 1991: 3-7